

An Aposteriorical Clusterability Criterion for k -Means and Simplicity of Clustering

Mieczysław A. Kłopotek

Institute of Computer Science of the Polish Academy of Sciences

ul. Jana Kazimierza 5, 01-248 Warszawa Poland

kłopotek@ipipan.waw.pl

April 25, 2017

Abstract

We define the notion of a well-clusterable data set from the point of view of the objective of k -means clustering algorithm and common sense.

The novelty introduced here is that one can a posteriori (after running k -means) check if the data set is well-clusterable or not.

1 Introduction

It is a commonly observed phenomenon that most practically used clustering algorithms (like k -means) have a high theoretical computational complexity (are NP-hard), but at the same time in many (though not all) practical applications they perform quite well (converge quickly enough) yielding more or less usable output. Apparently then, the data must have the property that some data sets are better clusterable than other.

Though a number of attempts have been made to capture formally the intuition behind clusterability, none of these efforts seems to have been successful, as Ben-David exhibits in [8] in depth. He points at two important shortcomings of current state-of-the-art research results: the clusterability cannot be checked prior to applying a potentially NP-hard clustering algorithm to the data, and beside this, known clusterability criteria impose strong separation constraints. Though a recent paper by Ackerman [1] partially eliminates these problems, but regrettably at the expense of introducing

user-defined parameters that do not seem to be intuitive (in terms of one's imagination about what well-clusterable data are.).

Therefore in this paper we try a different approach to defining what well clusterable data are. In particular we do not require that we have to say beforehand (before clustering) whether or not the data is well-clusterable. Instead we require that one shall be able to state aposterioriacally *whether or not* the data is well-clusterable according to well-clusterability criteria that were assumed.

Our contribution encompasses:

- Two brands of well-clusterability criteria for data to be clustered via k -means algorithm, that can be verified ex-post (both positively and negatively) without great computational burden.
- Demonstration, that the structure of well-clusterable data (according to these criteria) is easy to recover.
- Demonstration that if well-clusterable data structure (in that sense) was not discovered by k -means++, then there is no such structure in the data.
- Demonstration that large gaps between data clusters are not sufficient to ensure well-clustrability.

The structure of this paper is as follows: In Section 2 we recall the previous work on the topic. In Section 3 we show that large gaps are not sufficient for well-clusterability. In Section 4 we introduce the first version of well-clusterability concept and show that data well-clustered in this sense are easily learnable via k -means++. This concept has the drawback that no data points (outliers) can lie in wide areas between the clusters. Therefore in Section 7 we propose a core-based well-clusterability concept and show that data well-clustered in this sense are also easily learnable via k -means++. The concept of cluster core itself is introduced and investigated in Section 5 and a method determining proper gap size under these new conditions is derived in Section 6. In Section 8 we draw some conclusions from this research.

2 The problem of clusterability in the previous work

Intuitively the clusterability shall be a function taking a set of points and returning a real value saying how "strong" or "conclusive" is the clustering

structure of the data [2]. This intuition, however, turns out not to be formalized in a uniform way so that quite a large number of formal definitions have been proposed. Ackerman and Ben-David in [2] studied several of these notions. They concluded that across the various formalizations, two phenomena co-occur: on the one hand well-clusterable data sets (with high "clusterability" value) are computationally easy to cluster (in polynomial time), but on the other hand identification whether or not the data is well-clusterable is NP-hard.

Ben-David [8] performed an interesting investigation of the concepts of clusterability from the point of view of the capability of "not too complex" algorithms to discover the cluster structure, (negatively) verifying the working hypothesis that "Clustering is difficult only when it does not matter" (the *CDNM* thesis).

He considered the following notions of clusterability, present in the literature:

- *Perturbation Robustness* meaning that small perturbations of distances / positions in space of set elements do not result in a change of the optimal clustering for that data set. Two brands may be distinguished: additive [2] and multiplicative ones [9] (the limit of perturbation is upper-bounded either by an absolute value or by a coefficient).
- ϵ -*Separatedness* meaning that the cost of optimal clustering into k clusters is less than ϵ^2 times the cost of optimal clustering into $k-1$ clusters [13]
- (c, ϵ) -*Approximation- Stability* [6] meaning that if the cost function values of two partitions differ by the factor c , then the distance (in some space) between the partitions is at most ϵ . As Ben-David recalls, this implies the uniqueness of optimal solution.
- α -*Centre Stability* [5] meaning, for any centric clustering, that the distance of an element to its cluster centre is α times smaller than the distance to any other cluster centre under optimal clustering.
- $(1 + \alpha)$ *Weak Deletion Stability* [4] meaning that given an optimal cost function value OPT for k centric clusters, then the cost function of a clustering obtained by deleting one of the cluster centres and assigning elements of that cluster to one of the remaining clusters should be bigger than $(1 + \alpha) \cdot OPT$.

Under these notions of clusterability algorithms have been developed clustering the data nearly optimally in polynomial times, when some constraints are matched by the mentioned parameters.

However, these conditions seem to be rather extreme. For example, given the (c, ϵ) -Approximation- Stability [6], polynomial time clustering requires that, in the optimal clustering (beside its uniqueness), all but an ϵ -fraction of the elements, are 20 times closer to their own cluster centre than to every other cluster centre. ϵ -Separatedness requires that the distance to its own cluster centre must be at least 200 times closer than to every other cluster element [13]. And this is still insufficient if the clusters are not balanced. A ratio of 10^7 is deemed by these authors as sufficient. $(1 + \alpha)$ Weak Deletion Stability [4] demands distances to other clusters being $\log(k)$ times the "average radius" of the own cluster. The perturbational stability [2] induces exponential dependence on the sample size.

Anyway, we can draw a certain important conclusion from these concepts of clusterability mentioned above: People agree that a data set is well clusterable if each cluster is distant (widely separated) from the other clusters.

This idea occurs in many other clusterability concepts. Eptner et al. [10] considers the data as clusterable when the minimum between - cluster separation exceeds the maximum in - cluster distance (called elsewhere "perfect separation").¹ Balcan et al. [7] proposes to consider data as clusterable if each element is closer to all elements in its cluster than to all other data (called also "nice separation").² Interestingly, k -means reflects the Balcan concept "on average" that is each element average squared distance to elements of the same cluster is smaller than the minimum (over other clusters) averaged squared distance to elements of a different cluster.

Recently also Ackerman et al. [1] derived a method for testing clusterability of data based on the large gap assumption. They investigate the histogram of (all) mutual dissimilarities between data points. If there is no data structure, the distribution should be unimodal. If there are distant clusters, then there will occur one mode for short distances (within clusters) and at least one for long distances (between clusters). Hence, to detect clusterability, they apply tests of multimodality, namely the Dip [11] and Silverman [14] tests.

But the criterion of a sufficiently large gap between clusters is not reflected in various clustering function objectives, like for example k -means which may reach an optimum with poorly separated clusters in spite of the fact that there exists an alternative partition of data with a clear separation between clusters in the data, as we will demonstrate in Section 3. Also in Section 3 we will

¹It has been shown in the literature that under this notion of well-clusterability single link algorithm can detect clusters separated in such a way. It has also been shown that centre based algorithms like k -means may fail to detect such clusters.

²It has been shown in the literature that this notion of well-clusterability is hard to decide in a data set.

demonstrate, that multimodel distributions can be detected by Ackerman’s method even if there is no structure in the data.

Ben-David [8] raises a further important point that it is usually (in practically all above mentioned methods except [1], which has a flaw by itself) impossible to verify apriori if the data fulfils the clusterability criterion because the conditions refer either to all possible clusterings or to optimal clustering so that we do not have the possibility to verify whether or not the data set is clusterable, before one starts clustering (but usually computing the optimum is NP-hard).

In this paper, however, we would like to stress that the situation is even worse. Even at the termination of the clustering algorithm we are unable to say whether or not the clustered data set turned out to be well-clusterable. For example, the ϵ -Separatedness criterion requires that we know the nearly optimal solution for clustering into k and $k - 1$ elements. While we can usually get the upper approximations for the cost functions in both cases, we need actually the lower approximation for $k - 1$ in order to decide ex post if the data was well-clusterable, and hence whether or not we can say that we approximated the correct solution in some way. But we get it only for $k = 2$, hence for higher k the issue is not decidable.

The issue of ex-post decision on clusterability seems nevertheless to be simpler to solve than the apriorical one, therefore we will attack it in this paper. We are unaware that such an issue was even raised in the past. Though the criteria of [10] and [7] can clearly be applied ex post to see that in the resulting clustering the clusterability criteria hold, but these approaches lack the solving of the inverse issue: what if the clusterability criteria are not matched by the result clustering - is the data unclusterable? Could no other algorithm discover the clusterable structure?

One shall note at this point that the approach in [1] is different with this respect. Compared to methods requiring finding the optimum first, Ackerman’s approach seems to fulfil Ben-David requirement, that we can see if there is clusterability in the data before starting the clustering process as the clusterability method is computationally optimal as the computation of the histogram of dissimilarities is quadratic in sample size. But at an in-depth-investigation, the Ackerman’s clusterability determination method misses one important point: it requires a user-defined parameter and the user may or may not make the right guess. Furthermore, even if clusterability is decided by Ackerman’s tests, it is still uncertain if k -means algorithm will be willing to find such a clustering that fits Ackerman’s clusterability criterion. Beside this, as visible in Figure 1, one can easily find counterexamples to their concept of clusterability.

So in summary the issue of an aposteriorical determination if the data

were clusterable, remains an open issue.

Therefore it seems to be justified to restrict oneself to a problem as simple as possible in order to show that the issue is solvable at all. So in this paper we will limit ourselves to the issue of clusterability for the purposes of k -means algorithm.³ Furthermore we restrict ourselves to determine such cases when the clusterability is decidable "for sure".

The first problem to solve seems to be to get rid of the dependence on the undecidedness of optimality of the obtained solution.

But before proceeding let us recall the k -means cost function definition.

$$Q(\mathcal{C}) = \sum_{i=1}^m \sum_{j=1}^k u_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 = \sum_{j=1}^k \frac{1}{n_j} \sum_{\mathbf{x}_i, \mathbf{x}_l \in C_j} \|\mathbf{x}_i - \mathbf{x}_l\|^2 \quad (1)$$

for a dataset \mathbf{X} under some partition $\mathcal{C} = \{C_1, \dots, C_k\}$ into the predefined number k of clusters, $C_1 \cup \dots \cup C_k = \mathbf{X}$, where u_{ij} is an indicator of the membership of data point \mathbf{x}_i in the cluster C_j having the centre at $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$.

The k -means algorithm starts with some initial guess of the positions of $\boldsymbol{\mu}_j$ for $j = 1, \dots, k$ and then alternating two steps: cluster assignment and centre update till some convergence criterion is reached, e.g. no changes in cluster membership. The cluster assignment step updates u_{ij} values so that each element \mathbf{x}_i is assigned to a cluster represented by the closest $\boldsymbol{\mu}_j$. The centre update step uses the update formula $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$.

k -means++ algorithm is a special case of k -means where the initial guess of cluster centres proceeds as follows. $\boldsymbol{\mu}_1$ is set to be a data point uniformly sampled from \mathbf{X} . The subsequent cluster centres are data points picked from \mathbf{X} with probability proportional to the squared distance to the closest cluster centre chosen so far. For details check [3]. Note that the algorithm proposed by [13] differs from the k -means++ only by the non-uniform choice of the first cluster centre (the first pair of cluster centres should be distant, and the choice of this pair is proportional in probability to the squared distances between data elements).

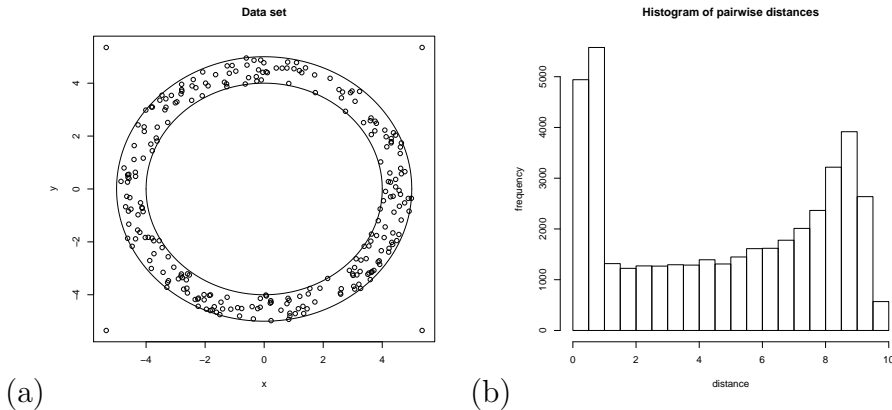


Figure 1: Illustration of a special case where Ackerman’s method [1] would falsely recognize clustering structure in the data (a) the data (b) the histogram of pairwise distances - two modes visible

3 Non-suitability of gap-based clusterability criteria for k -means

Let us discuss more closely the relationship between the gap-based well-clusterability concepts developed in the literature and the actual optimality criterion of k -means. Specifically let us consider the approaches to clusterability of [1], [10], [5] and [7].

Human intuition will tell us that if the groups of data points occur in the data and there are large spaces between these groups, then it should be these groups that will be chosen as the actual clustering. On the other hand if there are no gaps between the groups of data points, then one would expect that the data are not considered as well-clusterable. Furthermore, if the data is well-clusterable, one would expect a reasonable clustering algorithm to discover easily such a well-clusterable data structure.

However, these intuitions prove wrong in case of k -means.

Let us first point to the fact that [1] may indicate a clear bimodal structure in the data where there are no gaps in the data. We are unaware of anybody pointing at this weakness of well-clusterability in [1]: Imagine a thin ring uniformly covered with data points. We would be reluctant to say that there is a clustering structure in such data. Nonetheless we will see two obvious nodes in such data. See Figure 1. The thinner the ring, the more obvious

³ The k -means algorithm seems to be quite popular in various variants both in traditional, kernel and spectral clustering. Hence the results may be still of sufficiently broad importance.

the reason for the multimodality will be: we will get closer and closer to a cosine function.

On the other hand, even if there are gaps between groups of data, for example those required by [10], [5] or [7], k -means optimum may not lie in the partition exhibiting gap based well-clusterability property in spite of its existence, And not only for these gaps, but also for any arbitrary many times larger ones. As [10] is concerned, it may be considered as a special case of [7]. [5] may be viewed in turn as a strengthening of the concept of [10]. So let us discuss a situation in which both perfect and nice separation criteria are identical that is of two clusters. We will show that whatever α we assume in the α -stability concept, k -means fails to be optimal under unequal class cardinalities. Let these clusters, C_A, C_B be each enclosed in a ball of radius r and the distance between ball centres should be at least $4r$. We have demonstrated in [12] that under these circumstances the clustering of data into C_A, C_B reflects a local minimum of k -means cost function. But it is not the global minimum, as we will show subsequently. So at least for k -means the criteria of Epter and Balcan and Awasthi cannot be viewed as realistic definitions of well-clusterability.

For purposes of demonstration we assume that both clusters are of different cardinalities n_A, n_B and let $n_A > n_B$. We show that whatever distance between both clusters, we can get such a proportion of n_A/n_B that the clustering into C_A, C_B is not optimal.

Let us consider a d -dimensional space. Let us select the dimension that contributes most to the variance in cluster C_A . So the variance along this direction amounts to at least the overall variance divided by d . Let us denote this variance component as V_d . Consider this coordinate axis responsible for V_d to have the origin at the cluster centre of C_A . Project all the points of cluster C_A on this axis. The variance of projected points will be just V_d . Split the projected data set into two parts P_1, P_3 , one with coordinate above 0 and the rest. Let the centres of P_1, P_3 lie x_1, x_2 away from the cluster centre. Let there be n_1 data points of P_1 be at most x_1 distant from the origin, and n_2 more than x_1 from the origin. Let there be n_3 data points of P_3 be at most x_3 distant from the origin, and n_4 more than x_3 from the origin. Obviously, $n_1 + n_2 + n_3 + n_4 = n_A$. Under these circumstances, let us ask the question whether for a V_d some minimal values of x_1, x_3 are implied. Because if so, then by splitting the cluster C_A into P_1, P_3 and by increasing the cardinality of C_A the split into $P_1, P_2 \cup C_B$ will deliver a lower Q value so that for sure the clustering into C_A, C_B will be not be optimal.

So observe that

$$V_d \leq \left(\frac{x_1^2(n_1 + n_1^2/n_2 + n_1 + n_2)}{n_1 + n_2} \cdot (n_1 + n_2) + \frac{x_3^2 \cdot (n_3 + n_3^2/n_4 + n_3 + n_4)}{(n_3 + n_4)} \cdot (n_3 + n_4) \right) / n_A$$

that is

$$V_d \leq (x_1^2 \cdot (n_1 + n_1^2/n_2 + n_1 + n_2) + x_3^2 \cdot (n_3 + n_3^2/n_4 + n_3 + n_4)) / n_A$$

Note that we can delimit n_2, n_4 from below due to the relationship: $(r - x_1) \cdot n_2 \geq n_1 \cdot x_1$, $(r - x_3) \cdot n_4 \geq n_3 \cdot x_3$.

Therefore

$$V_d \leq (x_1^2 \cdot (n_1 + n_1^2 \cdot (r - x_1) / n_1 / x_1 + n_1 + n_2) + x_3^2 \cdot (n_3 + n_3^2 \cdot (r - x_3) / n_3 / x_3 + n_3 + n_4)) / n_A$$

Hence

$$V_d \leq (x_1^2 \cdot (2 \cdot n_1 + n_2) + n_1^2 \cdot (r - x_1) \cdot x_1 / n_1 + x_3^2 \cdot (2 \cdot n_3 + n_4) + n_3^2 \cdot (r - x_3) \cdot x_3 / n_3) / n_A$$

$$V_d \leq (x_1^2 \cdot (2 \cdot n_1 + n_2) + n_1 \cdot (r - x_1) \cdot x_1 + x_3^2 \cdot (2 \cdot n_3 + n_4) + n_3 \cdot (r - x_3) \cdot x_3) / n_A$$

$$V_d \leq (x_1^2 \cdot (n_1 + n_2) + n_1 \cdot r \cdot x_1 + x_3^2 \cdot (n_3 + n_4) + n_3 \cdot r \cdot x_3) / n_A$$

Let $n_1 + n_2$ be the minority among data points - then x_1 is larger and x_3 is smaller of the two.

$$V_d \leq (x_1^2 \cdot (n_1 + n_2) + n_1 \cdot r \cdot x_1 + (x_1 \cdot (n_1 + n_2) / (n_3 + n_4))^2 \cdot (n_3 + n_4) + n_3 \cdot r \cdot x_3) / n_A$$

$$V_d \leq (x_1^2 \cdot (n_1 + n_2) + n_1 \cdot r \cdot x_1 + x_1^2 \cdot (n_1 + n_2)^2 / (n_3 + n_4) + n_3 \cdot r \cdot x_3) / n_A$$

$$V_d \leq (x_1^2 \cdot (n_1 + n_2) \cdot n_A / (n_3 + n_4) + n_1 \cdot r \cdot x_1 + n_3 \cdot r \cdot x_3) / n_A$$

The above implies

$$V_d \leq (x_1^2 \cdot (n_1 + n_2) \cdot n_A / (n_3 + n_4) + (n_1 + n_2) \cdot r \cdot x_1 + (n_3 + n_4) \cdot r \cdot x_3) / n_A$$

Hence

$$V_d \leq ((x_3 \cdot (n_3 + n_4) / (n_1 + n_2))^2 \cdot (n_1 + n_2) \cdot n_A / (n_3 + n_4) + 2 \cdot (n_3 + n_4) \cdot r \cdot x_3) / n_A$$

We can delimit $n_1 + n_2$ from below due to relationship $x_3 \cdot (n_3 + n_4) \leq (n_1 + n_2) \cdot r$. Therefore

$$V_d \leq (x_3^2 \cdot (n_3 + n_4)^2 / (n_1 + n_2) \cdot n_A / (n_3 + n_4) + 2 \cdot (n_3 + n_4) \cdot r \cdot x_3) / n_A$$

$$V_d \leq (x_3^2 \cdot (n_3 + n_4)^2 \cdot r / x_3 / (n_3 + n_4) \cdot n_A / (n_3 + n_4) + 2 \cdot (n_3 + n_4) \cdot r \cdot x_3) / n_A$$

$$V_d \leq (x_3 \cdot r \cdot n_A + 2 \cdot (n_3 + n_4) \cdot r \cdot x_3) / n_A$$

So we obtain

$$V_d \leq 3 \cdot x_3 \cdot r$$

This means that

$$x_3 \geq V_d / 3 / r$$

what had to be shown in order to ensure that by scaling up n_A it pays off to split the first cluster and to attach the contents of the second one to one of the parts of the first, if we keep V_d when increasing n_A .

4 Our basic approach to clusterability

Let us stress at this point that the issue of well-clusterability is not only a theoretical issue, but it is of practical interest too. For example in creating synthetic data sets for investigating suitability of various clustering algorithms. But also after having performed the clustering process with whatever method we have, we need to answer one important question: whether or not the obtained clustering meets the expectation of the analyst.

These expectations may be divided into several categories:

- matching business goals,
- matching underlying algorithm assumptions,
- proximity to the optimal solutions.

Business goals of the clustering may be difficult to express in terms of data for an algorithm, or may not fit the algorithm domain or may be too expensive to collect prior to performing an approximate clustering.

For example, when one seeks a clustering that would enable efficient collection of cars to be scrapped (disassembly network), then one has to match multiple goals, like covering the whole country, maximum distance from client to the disassembly station, and of course the number of prospective clients, which is known with some degree of uncertainty. The distances to the clients are frequently not Euclidean in nature (due to geographical obstacles like rivers mountains etc.), while the preferred k -means algorithm works best with geometrical distances, no upper distance can be imposed etc. Other algorithms may induce same or different problems. So a posteriori one has to check if the obtained solution meets all criteria, does not violate constraints and is stable under fluctuation of the actual set of clients.

The other two problems are somehow related to one another. For example, you may have clustered the data being a subsample of the proper data set and the question may be raised how close the sub-sample cluster centres are to the cluster centres of the proper data set. Known methods allow to estimate this discrepancy given that we know that the cluster sizes do not differ too much. So prior to evaluating the correctness of cluster centre estimation we have to check if cluster proportions are within a required range (or if sub-sample size is relevant for such a verification). As another example consider methods of estimating closeness to optimal clustering solution under some general data distributions (like for the k -means++[3]), but the guarantees are quite loose. But at the same time the guarantees can be much tighter if the clusters are well-separated in some sense. So if we want to be sure with a reasonable probability that the obtained solution is sufficiently close to the optimum, we would need to check if the obtained clusters are well separated in the defined sense.

With this in mind, as mentioned, a number of researchers developed the concept of data clusterability. The notion of clusterability should intuitively reflect the following idea: if it is easy to see that there are clear-cut clusters in the data, then one would say that the data set is clusterable. "Easy to see" may mean either a visual inspection or some algorithm that quickly identifies the clusters. The well-established notion of clusterability would improve our understanding of the concept of the cluster itself - a well-defined clustering would be a clustering of clusterable points. This also would be a foundation for objective evaluation of clustering algorithms. The algorithm shall perform well for well-clusterable data and when the clusterability condition would be violated to some degree, the performance of a clustering algorithm is allowed to deteriorate also, but the algorithm quality would be measured on how the

clusterability violation impacts the deterioration of algorithm performance.

However, the issue turns out not to be that simple. As is well known, each algorithm seeking to discover a clustering may be betrayed somehow to fail to discover a clustering structure that is visible upon human inspection of data. So instead of trying to reflect human vision of clusterability of the data set independently of the algorithm, let us rather concentrate on finding a concept of clusterability that is both reflecting human perception and the minimum of cost function of a concrete algorithm, in our case k -means. We will particularly concentrate on its version called k -means++.

So let us define:

Definition 1. *A data set is well-clusterable with respect to k -means if (a) the data points may be split into subsets that are clearly separated by an appropriately chosen gap such that (b) the global minimum of k -means cost function coincides with this split and (c) with high probability the k -means++ algorithm discovers this split and (d) if the split was found, it may be verified that the data subsets are separated by the abovementioned gap and (e) if the k -means++ did not discover a split of the data fulfilling the requirement of the existence of the gap, then with high probability the split described by points (a) and (b) does not exist.*

In the paper [12] we have investigated conditions under which one can ensure that the minimum of k -means cost function is related to a clustering with (wide) gaps between clusters.

The conditions for clusterable data set therein are rather rigid, but serve the purpose of demonstration that it is possible to define properties of the data set that ensure this property of the minimum of k -means. Let us recall below the main result in this respect.

So assume that the data set encompassing n data points consists of k subsets such that each subset $i = 1, \dots, k$ can be enclosed in a ball of radius r_i . Let the gap (distance between surfaces of enclosing balls) between each pair of subsets amount to at least g , that is described below.

$$g \geq r_{\max} \sqrt{k \frac{M+n}{m}} \quad (2)$$

and

$$g \geq k r_{\max} \sqrt{n_p/2 + n_q/2 + n/2} \sqrt{\frac{2n}{n_p n_q}} \quad (3)$$

for any $p, q = 1, \dots, k; p \neq q$, when $n_i, i = 1, \dots, k$ is the cardinality of the cluster i , $M = \max_i n_i$, $m = \min_i n_i$,

Then it is claimed in that paper that the optimum of k -means objective is reached when splitting the data into the aforementioned subsets.

What are the implications? The most fundamental one is that the problem is decidable.

Theorem 1. (i) *If the data set is well-clusterable with a gap defined by formulas (2) and (3), then with high probability k -means++ (after an appropriately chosen number of repetitions) will discover the respective clustering.* (ii) *If k -means++ (after an appropriately chosen number of repetitions) does not discover a clustering matching formulas (2) and (3), then with high probability the data set is not well clusterable.*

The rest of the current section is devoted to the proof of the claims of this new theorem, proposed in the current paper.

If we obtained the split, then for each cluster we are able to compute the cluster centre, the radius of the ball containing all the data points of the cluster, and finally we can check if the gaps between the clusters meet the requirement of formulas (2) and (3). So we are able to decide that we have found that the data set is well-clusterable.

So let us look at the claim (i). As we already know, the global minimum of k -means coincides with the separation by abovementioned gaps. Hence if there exists a positive probability, that k -means++ discovers the appropriate split, then by repeating independent runs of k -means++ and picking the split minimising k -means cost function we will increase the probability of finding the global minimum. We will show that we know the number of repetitions needed in advance, if we assume the maximum value of the quotient M/m .

First consider the easiest case of all clusters being of equal sizes ($M = m$). Then the above equations can be reduced to ($r = r_{max}$)

$$g \geq r\sqrt{k(k+1)} \quad (4)$$

$$g \geq rk\sqrt{2k+k^2} \quad (5)$$

A diagram of dependence of g/r on k is depicted in Figure 2

Now let us turn to k -means++ seeding. If already i distinct clusters were seeded, then the probability that a new cluster will be seeded (under our assumptions) amounts to at least

$$\begin{aligned} \frac{(k-i)g^2}{(k-i)g^2 + ir^2} &\geq \frac{(k-i)r^2k^2(k+1)^2}{(k-i)r^2k^2(k+1)^2 + ir^2} \\ &= \frac{(k-i)k^2(k+1)^2}{(k-i)k^2(k+1)^2 + i} \geq \frac{k^2(k+1)^2}{k^2(k+1)^2 + (k-1)} \end{aligned}$$



Figure 2: Dependence of the gap g on k for clusters of equal radius and equal cardinalities.

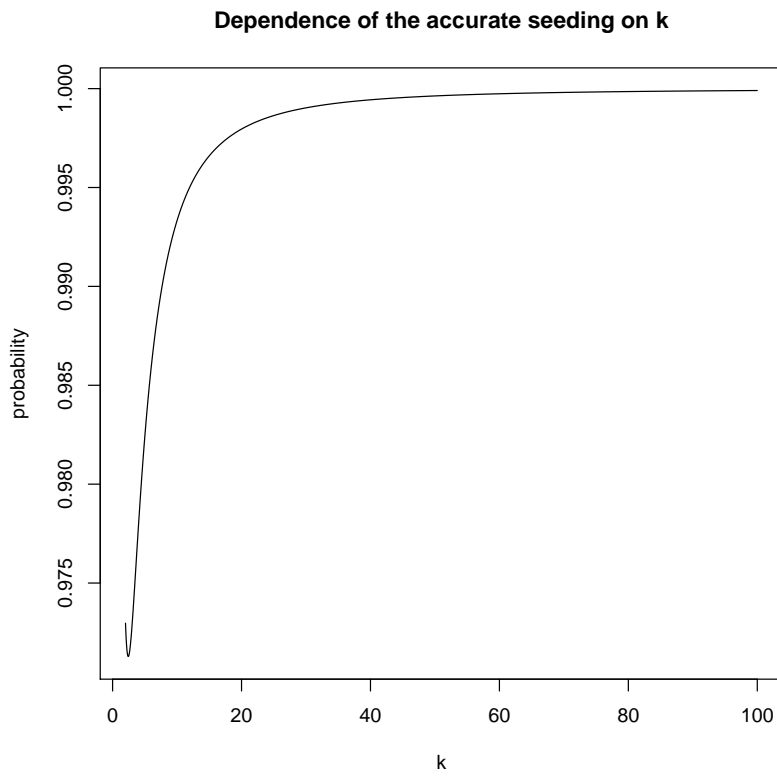


Figure 3: Probability of seeding each cluster on k for clusters of equal radius and equal cardinalities.

Hence the probability of accurate seeding amounts to

$$\left(\frac{k^2(k+1)^2}{k^2(k+1)^2 + (k-1)} \right)^{k-1}$$

The diagram of dependence of this expression on k is depicted in Figure 3.

Let us denote with Pr_{succ} the required probability of success in finding the global minimum. To ensure that the seeding was successful in Pr_{succ} (e.g. 95%) of cases, we need to rerun k -means++ at least R times, with R given by

$$\left(1 - \left(\frac{k^2(k+1)^2}{k^2(k+1)^2 + (k-1)} \right)^{k-1} \right)^R < 1 - Pr_{succ}$$

$$R \geq \frac{\log(1 - Pr_{succ})}{\log \left(1 - \left(\frac{k^2(k+1)^2}{k^2(k+1)^2 + (k-1)} \right)^{k-1} \right)}$$

But look at the following relationship:

$$\begin{aligned} & \left(\frac{k^2(k+1)^2}{k^2(k+1)^2 + (k-1)} \right)^{k-1} \\ &= \left(1 - \frac{k-1}{k^2(k+1)^2 + (k-1)} \right)^{k-1} \\ &= \left(1 - \frac{(k-1)^2}{k^2(k+1)^2 + (k-1)} \frac{1}{k-1} \right)^{k-1} \\ &\approx e^{-\left(\frac{(k-1)^2}{k^2(k+1)^2 + (k-1)} \right)} \end{aligned}$$

The exponent of the last expression approaches rapidly zero, so that with increasing k within a single pass of k -means++ the optimum is reached. In fact, already for $k=2$ we have an error of below 3%, for $k=8$, below 1%, for $k=30$ below 0.1%. See the Figure 3 for illustration.

Let us discuss clusters with same radius, but different cardinalities. Let m be the cluster minimum cardinality, and M respectively the maximum.

$$g \geq r \sqrt{k \frac{M+n}{m}} \tag{6}$$

$$g \geq kr \sqrt{\frac{n(n_p + n_q + n)}{n_p n_q}} \tag{7}$$

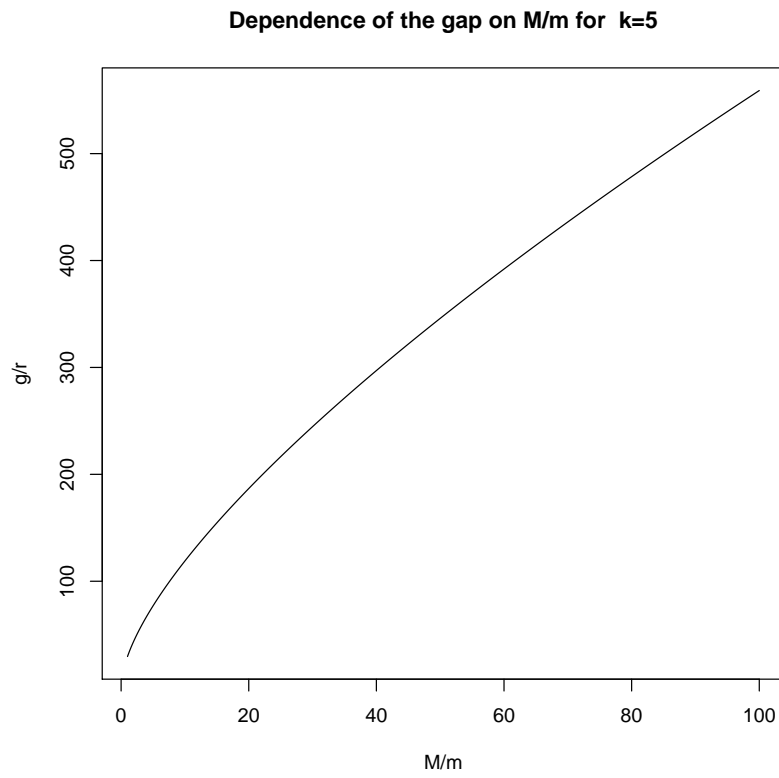


Figure 4: Dependence of the gap g for $k = 5$ for clusters of equal radius when varying cluster cardinalities.

for any $p, q = 1, \dots, k; p \neq q$, when $n_i, i = 1, \dots, k$ is the cardinality of the cluster i , $M = \max_i n_i$, $m = \min_i n_i$, Worst case g/r values are illustrated in Figure 4.

Now let us turn to k -means++ seeding. If already i distinct clusters were seeded, then the probability that a new cluster will be seeded (under our assumptions) amounts to at least

$$\begin{aligned} \frac{(k-i)mg^2}{(k-i)mg^2 + iMr^2} &\geq \frac{(k-i)mk^2n(1/m + 1/m + n/m^2)}{(k-i)mk^2n(1/m + 1/m + n/m^2) + iM} \\ &= \frac{(k-i)k^2n(2 + n/m)}{(k-i)k^2n(2 + n/m) + iM} \geq \frac{k^2n(2 + n/m)}{k^2n(2 + n/m) + (k-1)M} \end{aligned}$$

So again the probability of successful seeding will amount to at least:

$$\begin{aligned} &\left(\frac{k^2n(2 + n/m)}{k^2n(2 + n/m) + (k-1)M} \right)^{k-1} \\ &= \left(1 - \frac{(k-1)M}{k^2n(2 + n/m) + (k-1)M} \right)^{k-1} \\ &= \left(1 - \frac{(k-1)^2M}{k^2n(2 + n/m) + (k-1)M} \frac{1}{k-1} \right)^{k-1} \\ &\approx \exp \left(- \frac{(k-1)^2M}{k^2n(2 + n/m) + (k-1)M} \right) \end{aligned}$$

Even if M is 20 times as big as m , still the convergence to 1 is so rapid that already for $k = 2$ the clustering success is achieved with 95% success probability in a single repetition. An illustration is visible in Figure 5

So far we have concentrated on showing that if the data is well-clusterable, then within practically a single clustering run the seeding will have the property that each cluster obtains a single seed. But what about the rest of the run of k -means? As in all these cases $g \geq 2r$, then, as shown in [12], the cluster centres will never switch to balls encompassing other clusters, so that eventually the true cluster structure is detected and minimum of Q is reached. This would complete the proof of claim (i). The demonstration of claim (ii) is straight forward. If the data were well-clusterable then k -means++ would have failed to identify it with probability of at most $1 - Pr_{succ}$. As the well-clusterable data are in practice extremely rare, the failure of the algorithm to identify a well-clusterable structure induces with probability of at least Pr_{succ} that no such structure exists in the data.

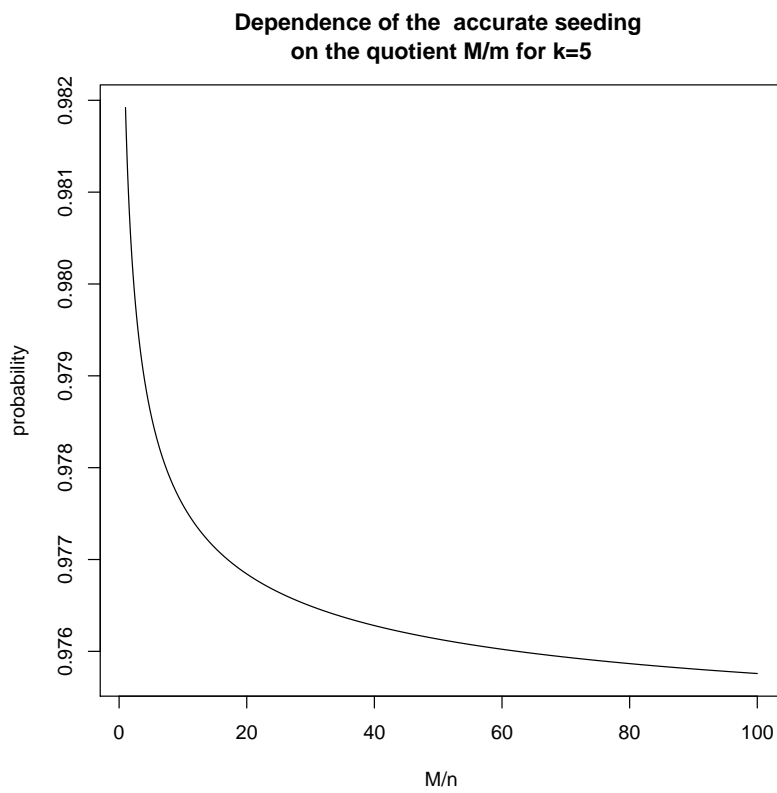


Figure 5: Probability of seeding each cluster for $k = 5$ for clusters of equal radius when varying cluster cardinalities.

5 Smaller gaps between clusters

In the previous section we considered well-clusterability under the assumption of large areas between clusters where no data points of any cluster will occur. Subsequently we show that this assumption may be relaxed so that spurious points are allowed between the major concentrations of cluster points. But to ensure that the presence of such points will not lead the k -means procedure astray, we will distinguish core parts of the clusters and will ensure by the subsequent theorem 3 that once a cluster core is hit by k -means initialisation procedure, the cluster is preserved over subsequent k -means iterations.

In [12] we have proven that

Theorem 2. *Let A, B be cluster centres. Let ρ_{AB} be the radius of a ball centred at A and enclosing its cluster and it also is the radius of a ball centred at B and enclosing its cluster. If the distance between the cluster centres A, B amounts to $2\rho_{AB} + g$, $g > 0$ (g being the "gap" between clusters), if we pick any two points, X from the cluster of A and Y from the cluster of B , and recluster both clusters around X and Y , then the new clusters will preserve the balls centred at A and B of radius $g/2$ (called subsequently "cores") each (X the core of A , Y the core of B).*

Here we shall demonstrate the validity of a complementary theorem.

Theorem 3. *Let A, B be cluster centres. Let ρ_{AB} be the radius of a ball centred at A and enclosing its cluster and it also is the radius of a ball centred at B and enclosing its cluster. Let ρ_{cAB} be the radius of a ball centred at A and enclosing "vast majority" of its cluster and it also is the radius of a ball centred at B and enclosing "vast majority" of its cluster. If the distance between the cluster centres A, B amounts to $2\rho_{AB} + g$, $g > 0$ ($g = 2\rho_{cAB}$ being the "gap" between clusters), if we pick any two points, X from the ball $B(A, \rho_{cAB})$ and Y from the ball $B(B, \rho_{cAB})$, and recluster both clusters around X and Y , then the new clusters will be identical to the original clusters around A and B .*

Definition 2. *If the gap between each pair of clusters fulfils the condition of either of the above two theorems, then we say that we have core-clustering.*

Proof. For the illustration of the proof see Figure 6.

The proof does not differ too much from the previous one and in fact the previous theorem is

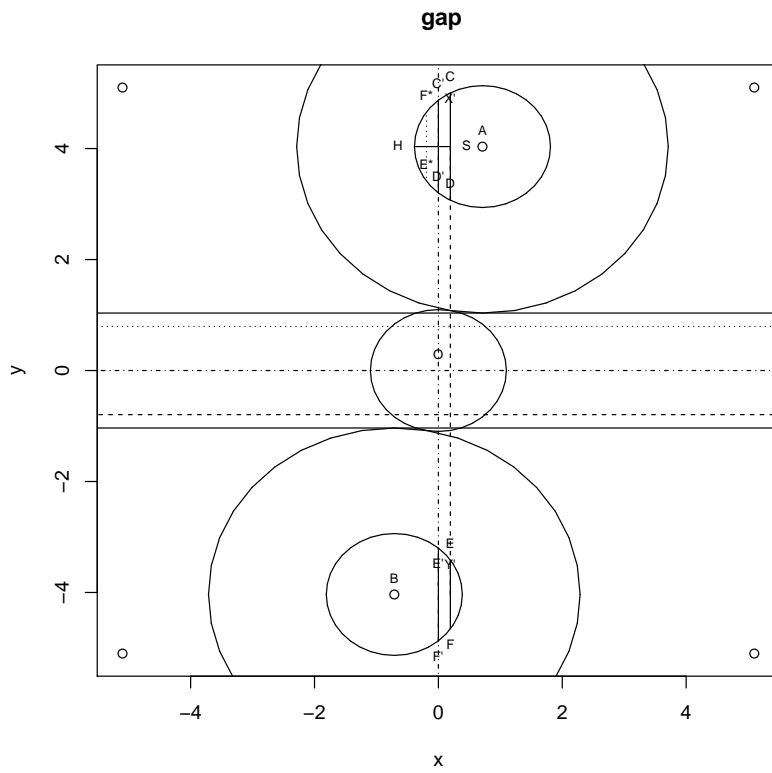


Figure 6: An illustrative figure for proof of the cluster preservation under a gap between cluster enclosing balls.

Consider the two points A, B being the two centres of double balls. The inner call represents the core of radius $r_{cAB} = g/2$, the outer ball of radius ρ ($\rho = \rho_{AB}$), enclosing the whole cluster. Consider two points, X, Y , one being in each core ball (presumably the cluster centres at some stage of the k -means algorithm). To represent their distances faithfully, we need at most a 3D space.

Let us consider the plane established by the line AB and parallel to the line XY . Let X' and Y' be orthogonal projections of X, Y onto this plane. Now let us establish that the hyperplane π orthogonal to X, Y , and passing through the middle of the line segment XY , that is the hyperplane containing the boundary between clusters centred at X and Y does not cut any of the balls centred at A and B . This hyperplane will be orthogonal to the plane of the Figure 6 and so it will manifest itself as an intersecting line l that should not cross outer circles around A and B , being projections of the respective balls. Let us draw two solid lines k, m between circles $O(A, \rho_{AB})$ and $O(B, \rho_{AB})$ tangential to each of them. Line l should lie between these lines, in which case the cluster centre will not jump to the other ball.

Let the line $X'Y'$ intersect with the circles $O(A, r_{cAB})$ and $O(B, r_{cAB})$ at points C, D, E, F as in the figure.

It is obvious that the line l would get closer to circle A , if the points X', Y' would lie closer to C and E , or closer to circle B if they would be closer to D and F .

Therefore, to show that it does not cut the circle $O(A, \rho)$ it is sufficient to consider $X' = C$ and $Y' = E$. (The case with ball $Ball(B, \rho)$ is symmetrical).

Let O be the centre of the line segment AB . Let us draw through this point a line parallel to CE that cuts the circles at points C', D', E' and F' . Now notice that centric symmetry through point O transforms the circles $O(A, r_{cAB}), O(B, r_{cAB})$ into one another, and point C' in F' and D' in E' . Let E^* and F^* be images of points E and F under this symmetry.

In order for the line l to lie between m and k , the middle point of the line segment CE shall lie between these lines.

Let us introduce a planar coordinate system centred at O with \mathcal{X} axis parallel to lines m, k , such that A has both coordinates non-negative, and B non-positive. Let us denote with α the angle between the lines AB and k . As we assume that the distance between A and B equals $2\rho + 2r_{cAB}$, then the distance between lines k and m amounts to $2((\rho + r_{cAB}) \sin(\alpha) - \rho)$. Hence the \mathcal{Y} coordinate of line k equals $((\rho + r_{cAB}) \sin(\alpha) - \rho)$.

So the \mathcal{Y} coordinate of the centre of line segment CE shall be not higher than this. Let us express this in vector calculus:

$$4(y_{OC} + y_{OE})/2 \leq ((\rho + r_{cAB}) \sin(\alpha) - \rho)$$

Note, however that

$$y_{OC} + y_{OE} = y_{OA} + y_{AC} + y_{OB} + y_{BE} = y_{AC} + y_{BE} = y_{AC} - y_{AE^*} = y_{AC} + y_{E^*A}$$

So let us examine the circle with centre at A. Note that the lines CD and E^*F^* are at the same distance from the line $C'D'$. Note also that the absolute values of direction coefficients of tangentials of circle A at C' and D' are identical. The more distant these lines are, as line CD gets closer to A, the y_{AC} gets bigger, and y_{E^*A} becomes smaller. But from the properties of the circle we see that y_{AC} increases at a decreasing rate, while y_{E^*A} decreases at an increasing rate. So the sum $y_{AC} + y_{E^*A}$ has the biggest value when C is identical with C' and we need hence to prove only that

$$(y_{AC'} + y_{D'A})/2 = y_{AC'} \leq ((\rho + r_{cAB}) \sin(\alpha) - \rho)$$

Let M denote the middle point of the line segment $C'D'$. As point A has the coordinates $((\rho + r_{cAB}) \cos(\alpha), (\rho + r_{cAB}) \sin(\alpha))$, the point M is at distance of $(\rho + r_{cAB}) \cos(\alpha)$ from A. But $C'M^2 = r_{cAB}^2 - ((\rho + r_{cAB}) \cos(\alpha))^2$.

So we need to show that

$$r_{cAB}^2 - ((\rho + r_{cAB}) \cos(\alpha))^2 \leq ((\rho + r_{cAB}) \sin(\alpha) - \rho)^2$$

In fact we get from the above

$$r_{cAB}^2 - ((\rho + r_{cAB}) \cos(\alpha))^2 \leq ((\rho + r_{cAB}) \sin(\alpha))^2 + \rho^2 - 2(\rho + r_{cAB})(\rho) \sin(\alpha)$$

$$r_{cAB}^2 \leq (\rho + r_{cAB})^2 + \rho^2 - 2(\rho + r_{cAB})(\rho) \sin(\alpha)$$

$$r_{cAB}^2 - \rho^2 \leq (\rho + r_{cAB})^2 - 2(\rho + r_{cAB})(\rho) \sin(\alpha)$$

$$(r_{cAB} - \rho)(r_{cAB} + \rho) \leq (\rho + r_{cAB})^2 - 2(\rho + r_{cAB})(\rho) \sin(\alpha)$$

$$(r_{cAB} - \rho) \leq (\rho + r_{cAB}) - 2\rho \sin(\alpha)$$

$$0 \leq 2\rho - 2\rho \sin(\alpha)$$

$$0 \leq 1 - \sin(\alpha)$$

which is obviously true, as \sin never exceeds 1. □

6 Core based global k -means minimum

In the paper [12] we have investigated conditions under which one can ensure that the minimum of k -means cost function is related to a clustering with (wide) gaps between clusters.

Based on the result of the preceding Section 5, we want to weaken these conditions requiring only that the big gaps exist between cluster cores and the clusters themselves are separated by much smaller gaps, equal to the size of the core.

In particular, let us consider the set of k clusters $\bar{\mathcal{C}} = \{\bar{C}_1, \dots, \bar{C}_k\}$ of cardinalities $\bar{n}_1, \dots, \bar{n}_k$ and with radii of balls enclosing the clusters (with centres located at cluster centres) $\bar{r}_1, \dots, \bar{r}_k$. Let each of these clusters \bar{C}_i have a core C_i of radius r_i and cardinality n_i around the cluster centre such that for $\mathfrak{p} \in [0, 1)$

$$Q(\{C_i\})/Q(\{\bar{C}_i\}) \geq 1 - \mathfrak{p}$$

We are interested in a gap g between cluster cores C_1, \dots, C_k such that it does not make sense to split each cluster \bar{C}_i into subclusters $\bar{C}_{i1}, \dots, \bar{C}_{ik}$ and to combine them into a set of new clusters $\mathcal{S} = \{S_1, \dots, S_k\}$ such that $S_j = \cup_{i=1}^k \bar{C}_{ij}$.

We seek a g such that the highest possible central sum of squares combined over the clusters \bar{C}_i would be lower than the lowest conceivable combined sums of squares around respective centres of clusters S_j . Let $Var(C)$ be the variance of the cluster C (average squared distance to cluster gravity centre; if referring to the core of the cluster, we still compute against the cluster centre, not the core centre, so also with the Q function). Let $C_{ij} = \bar{C}_{ij} \cap C_i$ be the core part of the subcluster \bar{C}_{ij} . Let r_{ij} be the distance of the centre of core subcluster C_{ij} to the centre of cluster \bar{C}_i . Let v_{ilj} be the distance of the centre of core subcluster C_{ij} to the centre of core subcluster C_{lj} . So the total k -means function for the set of clusters (C_1, \dots, C_k) will amount to:

$$Q(\bar{\mathcal{C}}) = \frac{1}{1 - \mathfrak{p}} Q(\mathcal{C}) = \frac{1}{1 - \mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k (n_{ij} Var(C_{ij}) + n_{ij} r_{ij}^2) \quad (8)$$

And the total k -means function for the set of clusters (S_1, \dots, S_k) will amount to:

$$Q(\mathcal{S}) \geq \sum_{j=1}^k \left(\left(\sum_{i=1}^k n_{ij} Var(C_{ij}) \right) + \left(\sum_{i=1}^k n_{ij} \right) \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij}}{\sum_{i=1}^k n_{ij}} \frac{n_{lj}}{\sum_{i=1}^k n_{ij}} v_{ilj}^2 \right) \right) \quad (9)$$

Should $(\overline{C}_1, \dots, \overline{C}_k)$ constitute the absolute minimum of the k -means target function, then $Q(\mathcal{S}) \geq Q(\overline{\mathcal{C}})$ should hold, which is fulfilled if :

$$\begin{aligned} \sum_{j=1}^k \left(\left(\sum_{i=1}^k n_{ij} \text{Var}(C_{ij}) \right) + \left(\sum_{i=1}^k n_{ij} \right) \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij}}{\sum_{i=1}^k n_{ij}} \frac{n_{lj}}{\sum_{i=1}^k n_{ij}} v_{ilj}^2 \right) \right) \\ \geq \frac{1}{1-\mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k (n_{ij} \text{Var}(C_{ij}) + n_{ij} r_{ij}^2) \end{aligned}$$

This implies:

$$\sum_{j=1}^k \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij} n_{lj}}{\sum_{i=1}^k n_{ij}} v_{ilj}^2 \right) \geq \frac{1}{1-\mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k (\mathfrak{p} n_{ij} \text{Var}(C_{ij}) + n_{ij} r_{ij}^2) \quad (10)$$

Note that $\text{Var}(C_{ij}) \leq r_{ij}^2$, so

$$\begin{aligned} \frac{1}{1-\mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k (\mathfrak{p} n_{ij} \text{Var}(C_{ij}) + n_{ij} r_{ij}^2) &\leq \frac{1}{1-\mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k (1+\mathfrak{p}) n_{ij} n_{ij} r_{ij}^2 \\ &= \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \sum_{i=1}^k \sum_{j=1}^k n_{ij} n_{ij} r_{ij}^2 \end{aligned} \quad (11)$$

To maximize $\sum_{j=1}^k n_{ij} r_{ij}^2$ for a single cluster C_i of enclosing ball radius r_i , note that you should set r_{ij} to r_i . Let $m_j = \arg \max_{j \in \{1, \dots, k\}} n_{ij}$. If we set $r_{ij} = r_i$ for all j except m_j , then the maximal r_{im_j} is delimited by the relation $\sum_{j=1; j \neq m_j}^k n_{ij} r_{ij} \geq n_{im_j} r_{im_j}$. So

$$\begin{aligned} \sum_{j=1}^k n_{ij} r_{ij}^2 &\leq \left(\sum_{j=1; j \neq m_j}^k n_{ij} \right) r_i^2 \min(2, (1 + \frac{\sum_{j=1; j \neq m_j}^k n_{ij}}{n_{im_j}})) \\ &\leq 2 \left(\sum_{j=1; j \neq m_j}^k n_{ij} \right) r_i^2 \end{aligned} \quad (12)$$

So if we can guarantee that the gap between cluster balls (of clusters from \mathcal{C}) amounts to g then surely

$$\sum_{j=1}^k \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij} n_{lj}}{\sum_{i=1}^k n_{ij}} v_{ilj}^2 \right) \geq g^2 \sum_{j=1}^k \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij} n_{lj}}{\sum_{i=1}^k n_{ij}} \right) \quad (13)$$

because in such case $g \leq v_{ilj}$ for all i, l, j .

By combining inequalities (10), (12) and (13) we see that the global minimum is granted if the following holds:

$$g^2 \sum_{j=1}^k \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij}n_{lj}}{\sum_{i=1}^k n_{ij}} \right) \geq 2 \frac{1+p}{1-p} \sum_{i=1}^k \left(\sum_{j=1; j \neq m_j}^k n_{ij} \right) r_i^2 \quad (14)$$

One can distinguish two cases: either (1) there exists a cluster S_t containing two subclusters C_{pt}, C_{qt} such that $t = \arg \max_j |C_{pj}|$ and $t = \arg \max_j |C_{qj}|$ (maximum cardinality subclasses of their respective original clusters C_p, C_q or (2) not.

Consider the first case. Let C_p, C_q be the two clusters where C_{pt} and C_{qt} be two subclusters of highest cardinality within C_p, C_q resp. This implies that $n_{pt} \geq \frac{1}{k}n_p, n_{qt} \geq \frac{1}{k}n_q$. Also this implies that for $i \neq p, i \neq q$ $n_{it} \leq n_i/2$.

$$\begin{aligned} & \sum_{j=1}^k \sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij}n_{lj}}{\sum_{i=1}^k n_{ij}} \\ & \geq \sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{it}n_{lt}}{\sum_{i=1}^k n_{it}} \\ & \geq \frac{n_{pt}n_{qt}}{\sum_{i=1}^k n_{it}} \\ & \geq \frac{n_{pt}n_{qt}}{n_p/2 + n_q/2 + \sum_{i=1}^k n_i/2} = \frac{n_{pt}n_{qt}}{n_p/2 + n_q/2 + n/2} \\ & \geq \frac{1}{k^2} \frac{n_p n_q}{n_p/2 + n_q/2 + n/2} \end{aligned}$$

Note that

$$2 \sum_{i=1}^k \left(\sum_{j=1; j \neq m_j}^k n_{ij} \right) r_i^2 \leq 2 \sum_{i=1}^k n_i r_i^2$$

So, in order to fulfil inequality (14), it is sufficient to require that

$$\begin{aligned} g & \geq \sqrt{\frac{2 \frac{1+p}{1-p} \sum_{i=1}^k n_i r_i^2}{\frac{1}{k^2} \frac{n_p n_q}{n_p/2 + n_q/2 + n/2}}} \\ & = k \sqrt{n_p/2 + n_q/2 + n/2} \sqrt{\frac{2 \frac{1+p}{1-p} \sum_{i=1}^k n_i r_i^2}{n_p n_q}} \\ & = k \sqrt{n_p + n_q + n} \sqrt{\frac{\frac{1+p}{1-p} \sum_{i=1}^k n_i r_i^2}{n_p n_q}} \quad (15) \end{aligned}$$

This of course maximized over all combinations of p, q .

Let us proceed to the second case. Here each cluster S_j contains a sub-cluster of maximum cardinality of a different cluster C_i . As the relation between S_j and C_i is unique, we can reindex S_j in such a way that actually C_j contains its maximum cardinality subcluster C_{jj} . Let us rewrite the inequality (14).

$$g^2 \sum_{j=1}^k \left(\sum_{i=1}^{k-1} \sum_{l=i+1}^k \frac{n_{ij}n_{lj}}{\sum_{i=1}^k n_{ij}} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k \left(\sum_{j=1; j \neq m_j}^k n_{ij} \right) r_i^2 \geq 0$$

This is met if

$$g^2 \sum_{j=1}^k \left(\sum_{i=1}^{j-1} \frac{n_{ij}n_{jj}}{\sum_{i=1}^k n_{ij}} + \sum_{l=j+1}^k \frac{n_{jj}n_{lj}}{\sum_{i=1}^k n_{ij}} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

This is the same as:

$$g^2 \sum_{j=1}^k \left(\sum_{i=1, \dots, j-1, j+1, \dots, k} \frac{n_{ij}n_{jj}}{\sum_{i=1}^k n_{ij}} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

This is fulfilled if:

$$g^2 \sum_{j=1}^k \left(\sum_{i=1, \dots, j-1, j+1, \dots, k} \frac{n_{ij}n_j/k}{n_j/2 + \sum_{i=1}^k n_i/2} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

Let M be the maximum over n_1, \dots, n_k . The above holds if

$$g^2 \sum_{j=1}^k \left(\sum_{i=1, \dots, j-1, j+1, \dots, k} \frac{n_{ij}n_j/k}{M/2 + n/2} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

Let m be the minimum over n_1, \dots, n_k . The above holds if

$$g^2 \sum_{j=1}^k \left(\sum_{i=1, \dots, j-1, j+1, \dots, k} \frac{n_{ij}m/k}{M/2 + n/2} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

This is the same as

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{j=1}^k \sum_{i=1, \dots, j-1, j+1, \dots, k} n_{ij} \right) - 2 \frac{1+p}{1-p} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{j=1}^k \left(\left(\sum_{i=1}^k n_{ij} \right) - n_{jj} \right) - 2 \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \right) \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\left(\sum_{j=1}^k \sum_{i=1}^k n_{ij} \right) - \left(\sum_{j=1}^k n_{jj} \right) - 2 \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \left(\sum_{i=1}^k (n_i - n_{ii}) r_i^2 \right) \right) \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\left(\sum_{i=1}^k n_i \right) - \left(\sum_{j=1}^k n_{jj} \right) - 2 \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \right) \geq 0$$

$$g^2 \frac{m/k}{M/2 + n/2} \left(\sum_{i=1}^k (n_i - n_{ii}) \right) - 2 \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \sum_{i=1}^k (n_i - n_{ii}) r_i^2 \geq 0$$

$$\sum_{i=1}^k (n_i - n_{ii}) \left(g^2 \frac{m/k}{M/2 + n/2} - 2 \frac{1+\mathfrak{p}}{1-\mathfrak{p}} r_i^2 \right) \geq 0$$

The above will hold, if for every $i = 1, \dots, k$

$$\begin{aligned} g &\geq r_i \sqrt{\frac{1+\mathfrak{p}}{1-\mathfrak{p}} \frac{2}{\frac{m/k}{M/2+n/2}}} \\ g &\geq r_i \sqrt{k \frac{1+\mathfrak{p}}{1-\mathfrak{p}} \frac{M+n}{m}} \end{aligned} \tag{16}$$

So the inequality (14) is fulfilled, if both inequality (15) and inequality (16) are held by an appropriately chosen g .

In summary we have shown that

Theorem 4. *Let $\overline{\mathcal{C}} = \{\overline{C}_1, \dots, \overline{C}_k\}$ be a partition of a data set into k clusters of cardinalities $\overline{n}_1, \dots, \overline{n}_k$ and with radii of balls enclosing the clusters (with centres located at cluster centres) $\overline{r}_1, \dots, \overline{r}_k$. Let each of these clusters \overline{C}_i have a core C_i of radius r_i and cardinality n_i around the cluster centre such that for $p \in [0, 1)$*

$$Q(\{C_i\})/Q(\{\overline{C}_i\}) \geq 1 - \mathfrak{p}$$

Then if the gap g between cluster cores C_1, \dots, C_k fulfils conditions expressed in formulas (15) and (16) then the partition $\overline{\mathcal{C}}$ coincides with the global minimum of the k -means const function for the data set.

7 Core based approach to clusterability

After the preceding preparatory work, we want to prove a theorem analogous to Theorem 1, but now allowing for smaller gaps between clusters.

Theorem 5. (i) *If the data set is well-clusterable with a gap defined by formulas (16) and (15), with r_i replaced by their maxima, then with high probability k -means++ (after an appropriately chosen number of repetitions) will discover the respective clustering. (ii) *If k -means++ (after an appropriately chosen number of repetitions) does not discover a clustering matching formulas (16) and (15) (with r_i replaced by their maxima), then with high probability the data set is not well clusterable.**

The rest of the current section is devoted to the proof of the claims of this theorem.

If we obtained the split, then for each cluster we are able to compute the cluster centre, the radius of the ball containing all the data points of the cluster but the \mathfrak{p} most distant ones, and finally we can check if the gaps between the cluster cores meet the requirement of formulas (16) and (15). So we are able to decide that we have found that the data set is well-clusterable.

So let us look at the claim (i). As we already know from preceding Section 6, the global minimum of k -means coincides with the separation by abovementioned gaps. Hence if there exists a positive probability, that k -means++ discovers the appropriate split, then by repeating independent runs of k -means++ and picking the split minimising k -means cost function we will increase the probability of finding the global minimum. We will show that we know the number of repetitions needed in advance, if we assume the maximum value of the quotient M/m .

We assume it is granted that

$$g \geq r \sqrt{k \frac{1 + \mathfrak{p}}{1 - \mathfrak{p}} \frac{M + n}{m}} \quad (17)$$

for any $i = 1, \dots, k$

$$g \geq kr \sqrt{\frac{1 + \mathfrak{p}}{1 - \mathfrak{p}} \frac{n(n_p + n_q + n)}{n_p n_q}} \quad (18)$$

for any $p, q = 1, \dots, k; p \neq q$, when $n_i, i = 1, \dots, k$ is the cardinality of the cluster i , $M = \max_i n_i$, $m = \min_i n_i$, For an illustration of this dependence see Figure 7.

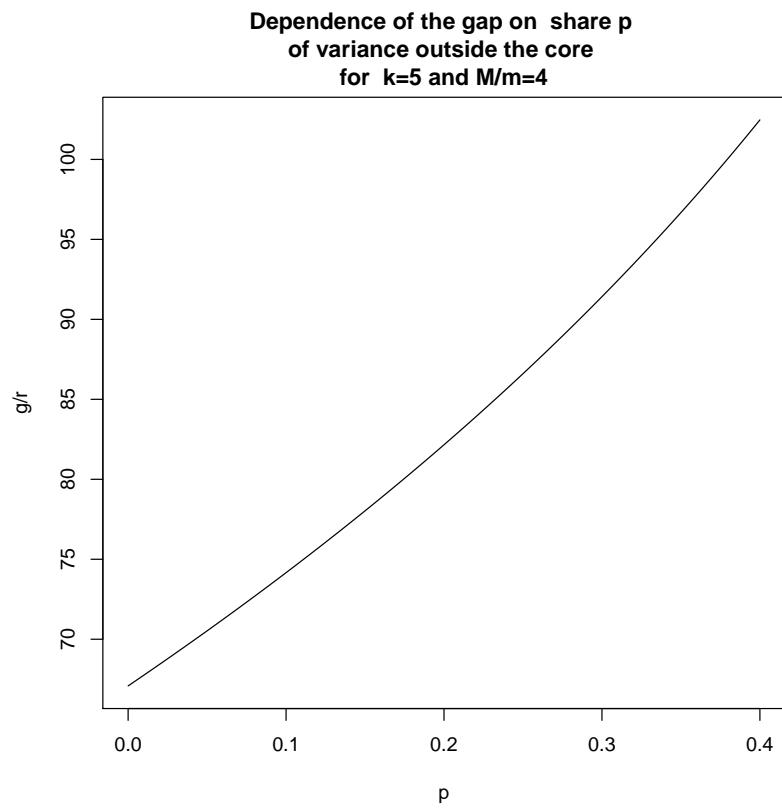


Figure 7: Dependence of g/r for $k = 5$ on the value of p

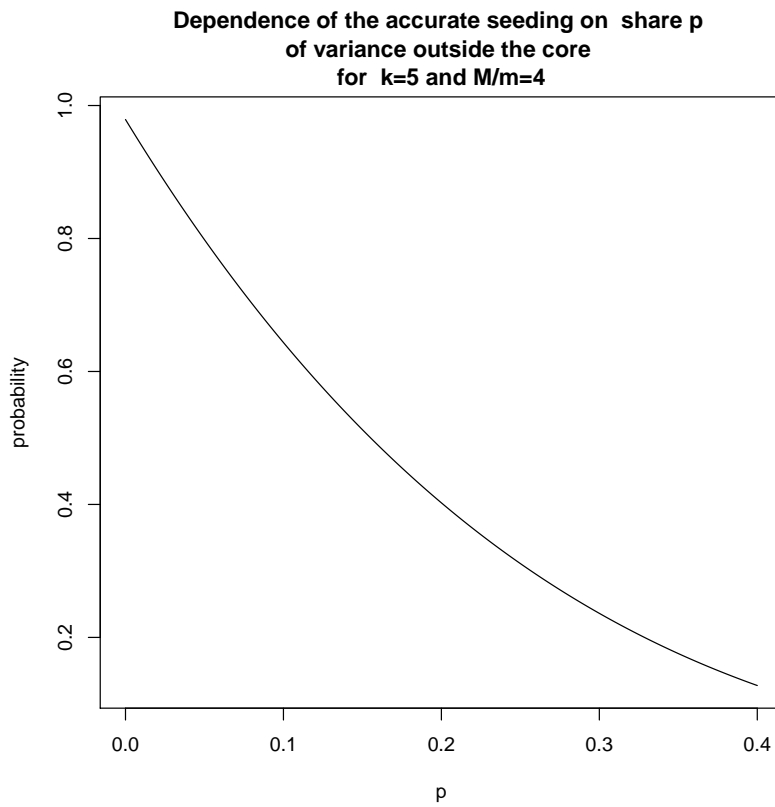


Figure 8: Probability of seeding each cluster for $k = 5$ for clusters of equal radius when varying cluster cardinalities.

So let us turn to k -means++ seeding. If already i distinct cluster cores were seeded, then the probability that a new cluster core will be seeded (under our assumptions) amounts to at least

$$\begin{aligned} \frac{(k-i)m(1-\mathfrak{p})g^2}{(k-i)mg^2 + iM\frac{1}{1-\mathfrak{p}}r^2} &\geq \frac{(k-i)mk^2(1-\mathfrak{p})\frac{1+\mathfrak{p}}{1-\mathfrak{p}}n(1/m + 1/m + n/m^2)}{(k-i)mk^2\frac{1+\mathfrak{p}}{1-\mathfrak{p}}n(1/m + 1/m + n/m^2) + iM\frac{1}{1-\mathfrak{p}}} \\ &= \frac{(k-i)k^2(1-\mathfrak{p})(1+\mathfrak{p})n(2+n/m)}{(k-i)k^2(1+\mathfrak{p})n(2+n/m) + iM} \\ &\geq \frac{k^2(1-\mathfrak{p})(1+\mathfrak{p})n(2+n/m)}{k^2(1+\mathfrak{p})n(2+n/m) + (k-1)M} \end{aligned}$$

So again the probability of successful seeding will amount to at least:

$$\begin{aligned} &\left(\frac{k^2(1-\mathfrak{p})(1+\mathfrak{p})n(2+n/m)}{k^2(1+\mathfrak{p})n(2+n/m) + (k-1)M} \right)^{k-1} \\ &= (1-\mathfrak{p})^{k-1} \left(1 - \frac{(k-1)M}{k^2(1+\mathfrak{p})n(2+n/m) + (k-1)M} \right)^{k-1} \\ &= (1-\mathfrak{p})^{k-1} \left(1 - \frac{(k-1)^2M}{k^2(1+\mathfrak{p})n(2+n/m) + (k-1)M} \frac{1}{k-1} \right)^{k-1} \\ &\approx (1-\mathfrak{p})^{k-1} \exp \left(- \frac{(k-1)^2M}{k^2(1+\mathfrak{p})n(2+n/m) + (k-1)M} \right) \end{aligned}$$

For an illustration of this dependence see Figure 8

Apparently in the limit the above expression lies at about $(1-\mathfrak{p})^{k-1}$.

So to achieve the identification of the clustering with probability of at least Pr_{succ} (e.g. 95%), we will need R runs of k -means++ where

$$R = \frac{\log(1 - Pr_{succ})}{\log(1 - (1-\mathfrak{p})^{k-1})}$$

Note that

$$1 - (1-\mathfrak{p})^{k-1} \approx 1 - e^{-\mathfrak{p}(k-1)} \approx 1 - e^{-\mathfrak{p}k}$$

The effect of doubling k is

$$\frac{1 - e^{-\mathfrak{p}2k}}{1 - e^{-\mathfrak{p}k}} = \frac{(1 - e^{-\mathfrak{p}k})(1 + e^{-\mathfrak{p}2k})}{1 - e^{-\mathfrak{p}k}} = 1 + e^{-\mathfrak{p}2k}$$

that is it is sublinear in the expression $1 - (1-\mathfrak{p})^{k-1}$, hence R grows slower than reciprocally logarithmically in k and p . For an illustration of this relation see Figure 9

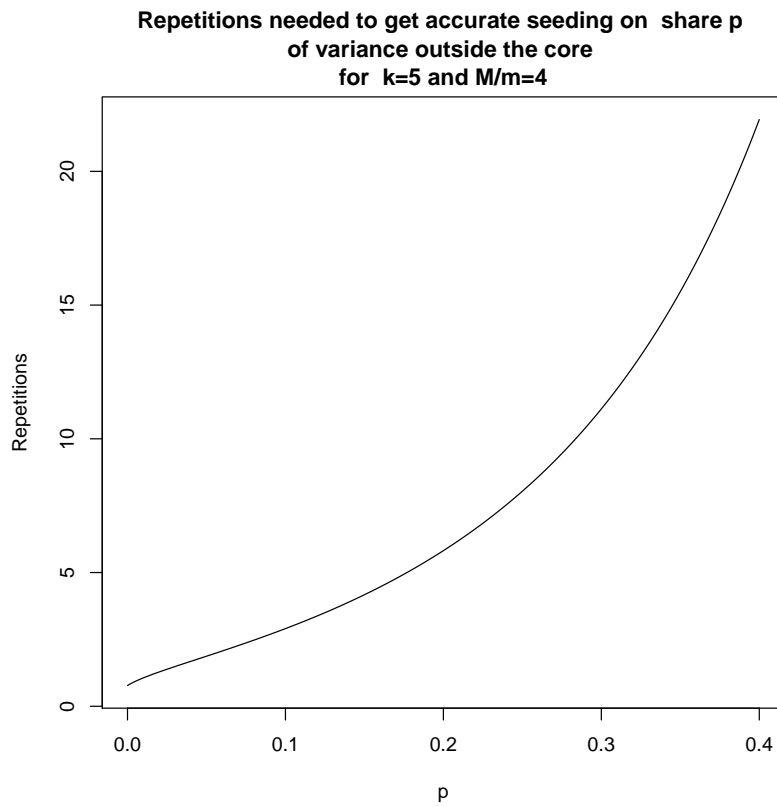


Figure 9: Probability of seeding each cluster for $k = 5$ for clusters of equal radius when varying cluster cardinalities.

So far we have concentrated on showing that if the data is well-clusterable, then within practically reasonable number of k -means++ runs the seeding will have the property that each cluster obtains a single seed. But what about the rest of the run of k -means? As shown in Section 5, the cluster centres will never switch to balls encompassing other clusters, so that eventually the true cluster structure is detected and minimum of Q is reached. This would complete the proof of claim (i). The demonstration of claim (ii) is straight forward. If the data were well-clusterable then k -means++ would have failed to identify it with probability of at most $1 - Pr_{succ}$. As the well-clusterable data are in practice extremely rare, the failure of the algorithm to identify a well-clusterable structure induces with probability of at least Pr_{succ} that no such structure exists in the data.

8 Conclusions

We have defined the notion of a well-clusterable data set from the point of view of the objective of k -means clustering algorithm and common sense in two variants - without any data points in the large gaps between clusters. The novelty introduced here, compared to other work on this topic, is that one can a posteriori (after running k -means) check if the data set is well-clusterable or not.

Let make a comparison to the results of other investigators in the realm of well-clusterability, in particular presented in [2, 13, 6, 5, 7, 1, 10].

If the data is well-clusterable according to criteria of Perturbation Robustness, or ϵ -Separatedness, or (c, ϵ) -Approximation- Stability or α -Centre Stability or $(1+\alpha)$ Weak Deletion Stability or Perfect Separation, one can reconstruct the well-clustered structure using appropriate algorithm. But only in case of Perfect Separation or Nice Separation, you can decide that you have found the structure, if you have found it. Note that you have no warranty that you will find Nice Separation if it is there. But for none of these ways of understanding well-clusterability we are able to decide (neither a priori nor a posteriori) that the data is not well-clusterable if the well-clustered structure was not found (unless by brute force).

The only exception constitutes formally the method of Multi-modality Detection, which tells us a priori that the data is or is not well-clusterable. However, as we have demonstrated, data can be easily found that can foolish this method, so that it discovers well-clusterability in case when there is none.

Under the definitions of well-clusterability presented in this paper, we get a completely new situation. It is guaranteed that if the well-clusterable structure is there, it will be detected with high probability. A posteriori one

can check that the structure found is the well-clusterable structure if it is so, with 100% certainty. Furthermore if the (k -means++) algorithm did not find a well-clusterable structure then with high probability it is not there in the data.

The paper contains a couple of other, minor contributions. The concept of cluster cores has been introduced such that if a seed of k -means once hits each core then there is guarantee that none of the cluster centres will ever leave the cluster. It has been shown that the number of reruns of k -means++ is small when a desired probability of success in finding the well-clusterability is being targeted. Numerical examples show that several orders of magnitude smaller gaps between clusters, compared to [13], are required in order to state that the data is well clusterable, and still the probability of detecting the well clusterable structure is much higher (even close to one in a single k -means++ run).

The procedure elaborated for constructing a well clusterable data set, ensuring that the k -means cost function absolute minimum is reached for a predefined data partition may find applications in some testing procedures of clustering algorithms.

Of course a number of research questions with respect to the topic of this paper remain open. First of all the issue of constructing tight (or at least tighter) bounds for estimation of required gaps between clusters. Second an investigation how the violations of these minimum values influence the capability of k -means algorithms to detect either the absolute minimum of their cost function or achieving a partition that is intuitively considered by humans as "good clustering".

References

- [1] Margareta Ackerman, Andreas Adolffson, and Naomi Brownstein. An effective and efficient approach for clusterability evaluation. *CoRR*, abs/1602.06687, 2016. earlier notions of clusterability 32,11,12,8,2,9 . 7,6.
- [2] Margareta Ackerman and Shai Ben-David. Clusterability: A theoretical study. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 1–8, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

- [3] D. Arthur and S. Vassilvitskii. k -means++: the advantages of careful seeding. In N. Bansal, K. Pruhs, and C. Stein, editors, *Proc. of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA 2007, pages 1027–1035, New Orleans, Louisiana, USA, 7-9 Jan. 2007. SIAM.
- [4] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k -median and k -means clustering. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 309–318, Washington, DC, USA, 2010. IEEE Computer Society.
- [5] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, January 2012.
- [6] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009*, pages 1068–1077, 2009.
- [7] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 671–680, New York, NY, USA, 2008. ACM.
- [8] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. <https://arxiv.org/abs/1501.00437>, 2015.
- [9] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Comb. Probab. Comput.*, 21(5):643–660, September 2012.
- [10] S. Epter, M. Krishnamoorthy, and M. Zaki. Clusterability detection and initial seed selection in large datasets. technical report 99-6, rensselaer polytechnic institute, computer science dept., rensselaer polytechnic institute, troy, ny 12180, 1999.
- [11] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Ann. Statist.*, 13(1):70–84, 1985.
- [12] R. Kłopotek and M. Kłopotek. On the discrepancy between kleinberg’s clustering axioms and k -means clustering algorithm behavior. <https://arxiv.org/abs/1702.04577>, 2017.

- [13] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28:1–28:22, January 2013. 0.0000001 is epsilon so that $\epsilon^2 \leq \text{target kmeans for } k / \text{target kmeans for } k-1$.
- [14] B.W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):97–99, 1981.